3.1 Experiment setup

3.1.1 Data collection

Numerous strategies considered in an attempt to procure data that belonged to the war veterans. Transcripts of podcasts and YouTube videos involving accounts of wars from the veterans, books that were written by the ex-servicemen, the public dataset that had diaries, and letters of first world war soldiers were a few sources. However, as an inference, all of these sources were highly specific to the negative aspects and impacts of war and eventually would add bias to the data.

On account of being a platform that is widely used by a large number of service-men and the civilians, Twitter was selected as the platform to extract data. Several verified Twitter pages linked to the US Army were manually analyzed and a few profiles of the veterans were used as an initial set. Fitting representation was ensured by selecting 6 veterans belonging to all the genders. There was also no divide on the number of Tweets. The initial set had Twitter profiles with as few as 57 tweets and as many as 39,000 tweets. The succeeding veterans were picked by scouring through the followers of the veterans in the original set based on some keywords like the army, us-army, military in their bio.

A collection of 20 veterans was used for the initial proof of concept. However, we intend to execute the study with 100 soldiers and 100 civilians.

Similarly, 6 civilians were selected with variations in age, sex, and the number of tweets. The profiles from the initial set were later parsed to get random followers. A collection of 20 commoners was used for the initial small-scale proof of concept.

3.1.2 Data process and analysis

3.1.2 Data Process and Analysis

Once the data is gathered and saved in the CSV format, we then start processing the gathered data where we come up with a set of results to prove our hypothesis. Sentiment analysis can be broadly categorized into two kinds, based on the type of output the analysis generates. Under our processing, we are trying to label text to be under “positive” and “negative”. Here we take into account the tweets of all the selected veterans from our data and then run an analysis to gather information of there sentiments braking then into two parts, that is world count of how much words they have used are under the world cloud of being ‘Happy’, and the same for the one under ‘Sad’. This identical analysis is then run on tweets by ordinary people. Finally, we will categorise these words to check the two sets of results which will ultimately help us identify an individual being in the state of depression with respects to the number of words count of “positive” and “negative” results.

The initial step was to make a sack of remarkable words. One significant perception that we went over was that not everyone of a kind word is significant for our handling. Thus, we focused on descriptive words which are profoundly educational of positive and negative assessments. The next steps that we took under processing our data were 1) Removing punctuation 2) Tokenizing 3) Removing Stop words 4) Generating tuple and 5) Finally, extracting the adjectives.

The first step is to remove punctuation followed by Tokenizing which is the way toward separating a goliath string into a rundown of words. “NLTK”, a python library is used for this process. “Stop words”, where pull-out, as they do not change any meaning of a sentence hence, can be ignored. Tuples were generated with each word and ts part of speech. Finally, the adjectives were extracted from the tweets of both the veterans and the civilians. And was then compared with the word cloud that was created.

Finally, we compared the count of the positive and negative words used by the two sets of people to come up with a result that displayed whether the veterans were targeted to depression and suicidal.

